

# EXACT MAXIMUM-ENTROPY ESTIMATION WITH FEYNMAN DIAGRAMS

TOMER M. SCHLANK, RAN J. TESSLER, AND AMITAI ZERNIK

ABSTRACT. A longstanding open problem in statistics is finding an explicit expression for the probability measure which maximizes entropy with respect to given constraints. In this paper a solution to this problem is found, using perturbative Feynman calculus. The explicit expression is given as a sum over weighted trees.

## CONTENTS

1. Introduction	1
1.1. Kullback-Liebler constraint problems	3
1.2. Moment trees	5
1.3. Cumulant trees	6
Acknowledgements	8
2. Feynmann Calculus: Perturbative Expression for Critical Point	8
2.1. Coordinate Expression for the Perturbed Critical Point	10
3. Proofs of the main theorems	12
3.1. Proof of moment tree theorems	13
3.2. Proofs of cumulant tree theorems	15
Appendix A. Proof of Lemma 3	16
References	17

## 1. INTRODUCTION

Given a finite set  $\Sigma$ , the relationship between distributions on  $\Sigma$  and an observable  $r : \Sigma \rightarrow \mathbb{R}$  on  $\Sigma$  is a two way street. First, given a distribution  $P_0$  on  $\Sigma$  we can ask for the expectation of  $r$ ,  $\mathbb{E}_{P_0} r$ .

In the other direction, suppose the distribution  $P_0$  is unknown but we are given, for  $1 \leq i \leq k$ , the expectation  $\rho_i$  of some observable  $r_i : \Sigma \rightarrow \mathbb{R}$  with respect to  $P_0$ . Of course, in general there will be infinitely many distributions  $P$  such that

$$(1) \quad \mathbb{E}_P r_i = \rho_i, \quad \forall 1 \leq i \leq k.$$

Arguably, among these distributions  $P$  the one that maximizes the entropy is the one that best reflects the information given by the expectations. This approach to estimation was first expounded by E. T. Jaynes in two papers in 1957 [Ja57I],[Ja57II].

A classical theorem of Ludwig E. Boltzmann shows that the maximum entropy distribution  $P$  belongs to a finite-dimensional exponential family of distributions parameterized by  $\lambda_i$ ,  $0 \leq i \leq k$ . It is not hard to prove that these parameters are analytic functions of  $\{\rho_i\}_{i=1}^k$ . It is possible to directly compute the first few orders of the series expansion of each  $\{\lambda_i\}_{i=0}^k$  in terms of  $\{\rho_i\}_{i=1}^k$ . For example, the quadratic approximation defines the *linear regression* multivariate normal distribution. As one tries to calculate higher orders, however, the computations quickly get out of hand. Many algorithms have been proposed for approximating the distribution numerically or by other means.

In this paper we give an explicit, combinatorial formula for computing the full Taylor expansion of  $\lambda_i(\rho_1, \dots, \rho_k)$  in terms of the joint moments of the observables  $r_i$  with respect to the uniform distribution. An alternative formula in terms of cumulants is also proven, which is computationally more efficient.

In estimation problems, one is often not interested in the distribution  $P$  itself, but rather in the expectation value  $\sigma := \mathbb{E}_P s$  of some observable  $s : \Sigma \rightarrow \mathbb{R}$ . In Theorems 14 and 9 we compute the Taylor expansion of  $\sigma(\rho_1, \dots, \rho_k)$ .

It is worth mentioning that given the higher moments of  $\{r_i\}$ , (or  $\{r_i\} \cup \{s\}$ ) the computation is completely independent of the size of the alphabet  $\Sigma$ . More precisely, to compute the Taylor expansion to any given order  $d$  only the joint moments of order  $\leq d+1$  are required.

We believe that the formulas presented here can find applications in estimation and classification problems, and have considerable theoretical value as well.

The main tool for proving these results is a formula for the Taylor expansion of a perturbed critical point. Folk theorems along these lines go back at least to Richard P. Feynman, and constitute the "classical", or "tree level" part of what is generally called the Feynman Calculus. We found the exposition in [Et02] very useful. At any rate, the discussion in Section 2 is completely self-contained, and readers may find it applicable to other optimization problems.

In the remainder of this section we state our main results. In section 2 we develop the main technical tool, Feynman calculus, and discuss its application to finding critical points of series expansions. This is used to prove the main results in Section 3.

**1.1. Kullback-Liebler constraint problems.** The problem of maximizing the entropy of a distribution  $P$  on an alphabet  $\Sigma$  subject to some constraints is an instance of the somewhat more general problem of *minimizing* the Kullback-Liebler divergence  $D(P||Q)$  of  $P$  relative to a given reference distribution  $Q$ . Indeed, if  $Q$  is the uniform distribution then  $D(P||Q) = \log |\Sigma| - H(P)$  where  $H(P)$  is the entropy of  $P$ . With this in mind, we can state our results as follows.

**Definition 1.** A *Kullback-Liebler constraint problem* (or KL constraint problem) consists of a triple  $(\Sigma, Q, \{r_i\}_{i=1}^k)$  where

- (1)  $\Sigma$  is a finite set called *the alphabet*.
- (2)  $Q = \{q_\sigma\}$  is a probability distribution on  $\Sigma$ , called *the reference probability distribution*.
- (3)  $r_i : \Sigma \rightarrow \mathbb{R}$  for  $1 \leq i \leq k$  are functions called the *constraint functions*.

We assume that  $q_\sigma > 0$  for all  $\sigma \in \Sigma$  and that the  $r_i$  are linearly independent as elements of the vector space  $\mathbb{R}^\Sigma$ ; one can reduce KL constraint problems which do not satisfy these conditions to ones that do, in an obvious way.

A KL constraint problem  $(\Sigma, Q, \{r_i\}_{i=1}^k)$  will be called *augmented* if we are given an additional *target function*  $s : \Sigma \rightarrow \mathbb{R}$ .

A KL constraint problem will be called *normalized* if  $\mathbb{E}_Q r_i = 0$  for  $1 \leq i \leq k$  and  $\mathbb{E}_Q r_i r_j = \delta_{ij}$  for  $1 \leq i, j \leq k$ .

*Remark 2.* There is no loss of generality in assuming that a KL problem is normalized. More precisely, given a KL constraint problem  $(\Sigma, Q, \{r_i\}_{i=1}^k)$ , one obtains a normalized problem in two steps.

- (i) Replace  $r_i$  by  $r'_i := r_i - \mathbb{E}_Q(r_i)$ , so  $\mathbb{E}_Q r'_i = 0$ .

Let  $E_0 \subset \mathbb{R}^\Sigma$  denote the sub-vector space  $\{f \in \mathbb{R}^\Sigma | \mathbb{E}_Q f = 0\}$ . Because  $q_\sigma > 0$  the covariance restricts to a non-degenerate positive definite pairing on  $E_0$ :

$$(2) \quad \text{Cov}(f, f') = \sum_{\sigma} q_{\sigma} \cdot f(\sigma) \cdot f'(\sigma).$$

- (ii) Use the Gram-Schmidt process to replace the sequence  $\{r'_i\}_{i=1}^k$  by a Cov-orthonormal sequence  $\{r''_i\}_{i=1}^k$  which has the same linear span.

$(\Sigma, Q, \{r''_i\}_{i=1}^k)$  defines an equivalent KL-constraint problem, in the sense that the set of distributions  $P$  satisfying Equation (1) is equal to the set of distributions  $P$  satisfying

$$(3) \quad \mathbb{E}_Q r''_i = \rho''_i$$

where  $(\rho''_i)_{i=1}^k$  is obtained from  $(\rho_i)_{i=1}^k$  by the affine-linear transformation effecting the change from  $r_i$  to  $r''_i$  computed above.

Given a normalized KL constraint problem  $(\Sigma, Q, \{r_i\}_{i=1}^k)$  and sufficiently small parameters  $\rho_i \in \mathbb{R}$ ,  $1 \leq i \leq k$ , we consider the probability distribution  $P = \{p_\sigma\}_{\sigma \in \Sigma}$  minimizing the Kullback-Liebler divergence of  $P$  relative to  $Q$ ,

$$D_{KL}(P||Q) = \mathbb{E}_P \log \left( \frac{P}{Q} \right) = \sum_{\sigma \in \Sigma} p_\sigma \log \left( \frac{p_\sigma}{q_\sigma} \right)$$

subject to the constraints  $\mathbb{E}_P r_i = \rho_i$ .

By Lagrange multipliers (see the beginning of Subsection 3.1) we have

$$(4) \quad p_\sigma = q_\sigma \exp \left( -1 - \lambda_0 - \sum_{i=1}^k \lambda_i r_i(\sigma) \right)$$

for some numbers  $\lambda_i = \lambda_i(\rho_1, \dots, \rho_k) \in \mathbb{R}$  which we call *the exponential parameters*,

In Appendix A we prove

**Lemma 3.** For  $1 \leq i \leq k$  the exponential parameter  $\lambda_i$  is an analytic function of  $\rho_1, \dots, \rho_k$ .

Now suppose the normalized KL-constraint problem is augmented by a target function  $s : \Sigma \rightarrow \mathbb{R}$ . We call  $\sigma = \sigma(\rho_1, \dots, \rho_k) = \mathbb{E}_P s$  the *target expectation*.

Lemma 3 has the following corollary,

**Corollary 4.** The target expectation  $\sigma(\rho_1, \dots, \rho_k)$  is an analytic function of  $\rho_1, \dots, \rho_k$

*Proof.* We have

$$\sigma(\rho_1, \dots, \rho_k) = \mathbb{E}_P s = \mu(\lambda_1(\rho_1, \dots, \rho_k), \dots, \lambda_k(\rho_1, \dots, \rho_k))$$

for

$$\mu(\lambda_1, \dots, \lambda_k) = \frac{\sum_{\sigma} s(\sigma) q_\sigma \exp \left( - \sum_{i=1}^k \lambda_i r_i(\sigma) \right)}{\sum_{\sigma} q_\sigma \exp \left( - \sum_{i=1}^k \lambda_i r_i(\sigma) \right)}$$

which is clearly analytic in a neighborhood of  $\lambda_1 = \dots = \lambda_k = 0$ . Since the composition of analytic functions is analytic, the result follows.  $\square$

The upshot is that the Taylor expansions of  $\lambda_i(\rho_1, \dots, \rho_k)$ ,  $1 \leq i \leq k$  and of  $\sigma(\rho_1, \dots, \rho_k)$  converge in an open polydisc around  $\rho_1 = \dots = \rho_k = 0$ . We refer to the power series expansions of these functions about the origin as *the perturbed exponential parameters*

$$(5) \quad \lambda_i(\rho_1, \dots, \rho_k) = \sum_I L_{i,I} \rho^I$$

and the perturbed expectation function

$$(6) \quad \mu(\rho_1, \dots, \rho_k) = \sum_I M_I \rho^I$$

here  $I = (i_1, \dots, i_k)$  ranges over  $\mathbb{Z}_{\geq 0}^k$  and  $\rho^I := \prod_{j=1}^k \rho_j^{i_j}$ .

**1.2. Moment trees.** Let  $(\Sigma, Q, \{r_i\}_{i=1}^k)$  be a normalized KL constraint problem. It will be notationally convenient to write  $r_0 : \Sigma \rightarrow \mathbb{R}$  for the constant function 1. If  $(\Sigma, Q, \{r_i\}_{i=1}^k, s)$  is an *augmented* KL constraint problem we introduce an additional function  $r_{k+1} : \Sigma \rightarrow \mathbb{R}$  which we set equal to the target function  $s$ .

**Definition 5.** A *rooted tree* (or RT)  $\Gamma$  is a tree with vertices  $V = V(\Gamma)$  and a nonempty set of edges  $E = E(\Gamma)$ , together with a distinguished leaf  $v_{out} \in V$ . We call the vertex to which the distinguished leaf  $v_{out}$  is connected the *root* of  $\Gamma$ . When  $|V(\Gamma)| > 2$  we define the *internal vertices*  $V_{in}$  to be the vertices which are not leaves. When  $|V(\Gamma)| = 2$  we set  $V_{in} = V \setminus \{v_{out}\}$ .

**Definition 6.** For  $j = 0, \dots, k+1$  a *j-moment tree*  $\Gamma$  for the normalized KL constraint problem  $(\Sigma, Q, \{r_i\}_{i=1}^k)$  is a rooted tree whose edges are labeled by  $\{r_i\}_{i=0}^{k+1}$  subject to the following conditions:

- Shape conditions: each vertex which is not a leaf has valency  $\geq 3$ .
- Labeling conditions: The edge connected to  $v_{out}$  is labeled by  $r_j$ ; the edges connected to the other leaves are labeled by  $r_1, \dots, r_k$ . All other edges are labeled by  $r_0, \dots, r_k$ .

A *moment tree* is a  $j$ -moment tree for some  $0 \leq j \leq k+1$ .

To a moment tree  $\Gamma$  we associate a *leaf multi-index*  $I_\Gamma = (a_1, \dots, a_k)$  with  $a_i$  the number of leaves with edges labeled by  $r_i$ . Note that since the valency of inner nodes is  $\geq 3$  there are only a finite number of moment trees with any given leaf multi-index.

**Definition 7.** Let  $\Gamma$  be a moment tree. Let  $v$  be an inner vertex of  $\Gamma$  of degree  $d$  with edges labeled by  $r_{i_1}, \dots, r_{i_d}$ .

(1) The *coupling value*  $C_v$  of  $v$  is

$$(7) \quad C_v = (-1)^{d+1} \mathbb{E}_Q r_{i_1} \cdots r_{i_d}.$$

(2) The *amplitude*  $A_\Gamma$  of  $\Gamma$  is

$$A_\Gamma = \prod_{v \in V_{in}(\Gamma)} C_v$$

**Theorem 8.** *Given a normalized KL constraint problem, the coefficients of the perturbed exponential parameters of Equation 5 are given by*

$$(8) \quad L_{j,I} = \sum_{\Gamma} \frac{A_{\Gamma}}{|Aut(\Gamma)|}$$

for  $1 \leq j \leq k$  where the sum runs over representatives  $\Gamma$  for all isomorphism types of  $j$ -moment trees  $\Gamma$ , and where  $Aut(\Gamma)$  is the group of automorphisms of  $\Gamma$ , i.e. the group of rooted tree automorphisms, preserving the labels and the root. with  $I_{\Gamma} = I$ .

**Theorem 9.** *Given an augmented normalized KL constraint problem, the perturbed expectation coefficients of Equation 6 are given by*

$$M_I = \sum_{\Gamma} \frac{A_{\Gamma}}{|Aut(\Gamma)|}$$

where the sum runs over representatives  $\Gamma$  for all isomorphism types of  $(k+1)$ -moment trees  $\Gamma$  with  $I_{\Gamma} = I$ .

*Remark 10.* Define the order of a tree  $\Gamma$  to be  $\sum_{i=1}^k a_i$  for  $(a_1, \dots, a_k) = I_{\Gamma}$  the leaf-multiindex associated with  $\Gamma$ . It is not hard to see that the order one contributions to the perturbed expectation give  $\mathbb{E}_Q s$ , and the order  $\leq 2$  contributions give the linear regression estimation of  $s$ , given the expectation and covariance values. In this sense Theorem 9 can be seen as a “perturbed regression” method, which generalizes linear regression to incorporate higher order moments.

**1.3. Cumulant trees.** Let  $(\Sigma, Q, \{r_i\}_{i=1}^k)$  be a normalized KL constraint problem. Now we will not have use for  $r_0$ ; but as before, if the problem is augmented we will denote  $r_{k+1} = s$ .

**Definition 11.** For  $j = 1, \dots, k+1$  a  $j$ -cumulant tree is a rooted tree whose edges are labeled by  $\{r_i\}_{i=1}^{k+1}$  subject to the following conditions:

- (1) Shape conditions: each vertex which is not a leaf has valency  $\geq 3$ .
- (2) Labeling conditions: The edge connected to  $v_{out}$  is labeled by  $r_j$ , all other edges (including edges connected to non-root leaves) are labeled by  $r_1, \dots, r_k$ .

A *cumulant tree* is a  $j$ -cumulant tree for some  $1 \leq j \leq k+1$ .

The *leaf multi-index* of  $\Gamma$  is  $I_{\Gamma} = (a_1, \dots, a_k)$  where  $a_i$  is the number of leaves whose edges are labeled by  $r_i$ . Again there are only a finite number of cumulant trees with any given multi-index.

**Definition 12.** Let  $\Gamma$  be a cumulant tree. Let  $v$  be an inner vertex of  $\Gamma$  of valency  $k$  whose edges are labeled by  $r_{i_1}, \dots, r_{i_d}$ .

(1) The *coupling value*  $C_v$  of  $v$  is

$$C_v = (-1)^{d+1} \kappa(r_{i_1}, \dots, r_{i_d})$$

where  $\kappa(r_{i_1}, \dots, r_{i_k})$  is the *joint cumulant* (cf. Eq (32)) of  $r_{i_1}, \dots, r_{i_k}$ .

(2) The *amplitude* of  $\Gamma$  is

$$(9) \quad A_\Gamma := \prod_{v \in V_{in}(\Gamma)} C_v$$

**Theorem 13.** *Given a normalized KL constraint problem, the coefficients of the perturbed exponential parameters, defined in 5, are given by*

$$(10) \quad L_{j,I} = \sum \frac{A_\Gamma}{|Aut(\Gamma)|} \text{ for } 1 \leq j \leq k,$$

where the sum runs over all isomorphism types of  $j$ -cumulant trees  $\Gamma$  with  $I_\Gamma = I$ .

**Theorem 14.** *Given an augmented normalized KL constraint problem, the coefficients of perturbed expectation, defined in 6, are given by*

$$M_I = \sum \frac{A_\Gamma}{|Aut(\Gamma)|},$$

where the sum runs over all isomorphism types of  $(k+1)$ -cumulant trees  $\Gamma$  with  $I_\Gamma = I$ .

These theorems are proved in Subsection 3.2.

*Remark 15.* The number of labeled trees is exponential where the base depends on the number of labels. Since cumulant trees have one less labels, working with them produces an exponential reduction in the complexity. In practice the improvement may be even more significant since cumulants tend to decay more rapidly than moments.

*Remark 16.* Consider the equivalence relation on moment trees which is generated by contracting edges labeled by 0. Equivalence classes of this relation correspond to cumulant trees, in an obvious way. This correspondence preserves the output edge label and the leaf multiindex, and it is possible to show that the contribution of each cumulant tree in Eq (10) is the sum of the contributions to Eq (8) of the moment trees in the corresponding equivalence class. A similar statement holds for the pertrubed expectation.

**Acknowledgements.** We thank O. Bozo, B. Gomberg, R.S. Melzer, A. Moscovitch-Eiger, R. Schweiger and D. Zernik for discussions related to the work presented here.

R.T. was partially supported by Dr. Max Rössler, the Walter Haefner Foundation and the ETH Zurich Foundation.

## 2. FEYNMANN CALCULUS: PERTURBATIVE EXPRESSION FOR CRITICAL POINT

We now introduce the main technical tool of this paper - using Feynman calculus to express critical points of functions.

Let  $\tau^x(y) : \mathbb{C}^{r+m} \rightarrow \mathbb{C}$  be an analytic function which we think of as a family of functions  $\tau^x : \mathbb{C}^m \rightarrow \mathbb{C}$  in  $y \in \mathbb{C}^m$ , parameterized by  $x \in \mathbb{C}^r$ . We assume  $\tau_0$  has a critical point at  $y = 0$  which is non-degenerate, i.e.  $\partial_y^2 \tau^x|_{x=0, y=0}$  defines a non-degenerate pairing on  $V = \mathbb{C}^m$ . For sufficiently small values of  $x$  and  $y$ ,  $\tau^x$  obtains a unique critical point  $y = \text{crit}(\tau^x)$ . In fact,  $x \mapsto \text{crit}(\tau^x)$  is an analytic function, and we will see that it admits an asymptotic expansion in terms of summation over trees. The entire discussion can be viewed as a kind of effective version of the contraction mapping proof of the implicit function theorem. We now explain this in more detail.

Let  $f(x, y) = \partial_y \tau^x : \mathbb{C}^r \times V \rightarrow V^*$  denote the partial derivatives of  $\tau^x$  in the  $y$  directions, where  $V^* = \text{Hom}_{\mathbf{Vect}_{\mathbb{C}}}(V, \mathbb{C})$ . Since  $\tau^0$  has a critical point at 0, we have  $f(0, 0) = 0$ , and  $f(x, y) = 0$  if and only if  $y$  is a critical point for  $\tau^x$ . Non-degeneracy of the critical point at  $y = 0$  amounts to saying  $B := \partial_y f|_{x=0, y=0} : V \rightarrow V^*$  is an invertible linear transformation

Define  $g : \mathbb{C}^r \times V \rightarrow V$  by

$$(11) \quad g^x(y) = y - B^{-1} \circ f(x, y)$$

Clearly,  $f(x, y) = 0$  iff  $g^x(y) = y$ . We have  $g^0(0) = 0$  and  $\partial_y g|_{x=0, y=0} = 0$ . By continuity we may find open neighborhoods  $0 \in U \subseteq \mathbb{C}^r, 0 \in W \subseteq V$  such that for any  $x \in U$ ,  $g^x(W) \subseteq W$  and  $d_y g^x|_W$  is a contraction, so that  $g^x|_W$  is a contraction.

Now by the Banach fixed point theorem, for any  $x \in U$  there exists a unique fixed point  $\text{crit}(\tau^x) \in W$  with  $g^x(\text{crit}(\tau^x)) = \text{crit}(\tau^x)$ . Moreover, for any  $x \in U$  the sequence  $y_n^x$  defined by  $y_0^x = 0, y_1^x = g^x(0), y_2^x = g^x(g^x(0)), \dots, y_{n+1}^x = g^x(y_n^x)$ , converges to  $\text{crit}(\tau^x)$ .

By the standard contraction-mapping proof of the implicit function theorem, one shows that the assignment  $x \mapsto \text{crit}(\tau^x)$  is an analytic



function  $U \rightarrow W$ . Write

$$\tau^x(y) = T_0^x + T_1^x y + \frac{1}{2!}(B + T_2^x)(y, y) + \sum_{l \geq 3} \frac{1}{l!} T_l^x(y, \dots, y),$$

where  $T_l^x \in \text{Sym}^l(V^*) \subset (V^*)^{\otimes l}$  are symmetric tensors,  $T_2^0 = 0$ . Note that we have natural identifications

$$(V^*)^{\otimes l} = (V^*)^{\otimes l-1} \otimes V^* = (V^{\otimes(l-1)})^* \otimes V^* = \text{Hom}_{\mathbf{Vect}_{\mathbb{C}}}(V^{\otimes(l-1)}, V^*)$$

and by abuse of notation we identify  $B : V \rightarrow V^*$  with the pairing  $B \in V^* \otimes V^*$ . In the same spirit, we consider  $T_l^x$  as a linear map  $V^{\otimes(l-1)} \rightarrow V^*$ . Unwinding definitions we see that

$$(12) \quad g^x(y) = -B^{-1}T_1^x - B^{-1}T_2^x(y) - \sum_{l \geq 3} \frac{1}{(l-1)!} B^{-1}T_l^x(y^{\otimes(l-1)})$$

We now explain how  $y_n^x := (g^x)^n(0)$  can be interpreted as a sum over rooted planar trees.

**Definition 17.** Let  $\Lambda$  be a rooted tree. The *height* of  $\Lambda$ ,  $h$ , is the length of the longest simple path  $(v_0, v_1, \dots, v_h)$  in  $\Lambda$  with  $v_0 = v_{out}$ . A *planar rooted tree* (or PRT)  $\Lambda$  is a rooted tree with an additional specification of a cyclic order on the set of edges incident to  $v$  for every  $v \in V$ .

Let  $\Lambda$  be a planar rooted tree. Fix some  $v \in V$ . Let  $\Lambda - v$  denote the graph obtained by removing  $v$  and all the edges incident to it from  $\Lambda$ . A subtree of  $v$  is any connected component of  $\Lambda - v$  which does not contain the distinguished leaf. Note that for  $\Lambda$  planar the subtrees of  $v$  are naturally ordered. Note that if  $\Lambda$  has height  $\leq h$  then the subtrees of the root are of height  $\leq h-1$  and completely specify the isomorphism type of  $\Lambda$ . This leads us to the following recursive construction.

For  $h \geq 0$  we define the set  $\mathcal{PRT}^{\leq h}$  of *representative planar rooted trees of height  $\leq h$*  by setting  $\mathcal{PRT}^{\leq 0} = \emptyset$  and  $\mathcal{PRT}^{\leq h} = \text{Seq}(\mathcal{PRT}^{\leq h-1})$  for  $h \geq 1$  where  $\text{Seq}(A) = \bigcup_{k \in \mathbb{Z}_{\geq 0}} A^{\times k}$  denotes the set of all finite sequences  $(a_1, \dots, a_k)$  with  $k \geq 0$  and  $a_i \in A$ . The set  $\mathcal{PRT} := \bigcup_{h=0}^{\infty} \mathcal{PRT}^{\leq h}$  is then a set of representatives for the isomorphism types of planar rooted trees. For  $l \geq 0$  define  $S_l^x : V^{\otimes l} \rightarrow V$  by  $S_l = B^{-1}T_{l+1}^x$ . Next we define the map  $\text{Cont}^x : \mathcal{PRT} \times \mathbb{C}^r \rightarrow V$  recursively on the height. For  $\mathcal{PRT}^{\leq 0}$  define it to be 0. For  $h \geq 1$  we define  $\text{Cont}^x(a_1, \dots, a_l)$  for  $(a_1, \dots, a_l) \in \mathcal{PRT}^{\leq h}$  by

$$\text{Cont}^x((a_1, \dots, a_l)) = -\frac{1}{l!} S_l^x \left( \bigotimes_{i=1}^s \text{Cont}^x(a_i) \right)$$

*Remark 18.* Somewhat less formally, given a PRT  $\Lambda$  we can compute  $\text{Cont}^x(\Lambda)$  by placing  $B^{-1} \in V \otimes V$  on the edges  $E(\Lambda)$  and  $\frac{(-1)}{(val(v)-1)!} \cdot T_{val(v)}^x \in (V^*)^{\otimes val(v)}$  on each vertex  $v \in V(\Lambda) \setminus \{v_{out}\}$  of valency  $val(v)$ , and then using the incidence pairing to contract the tensors. The result is  $\text{Cont}^x(\Lambda) \in V$ .

**Lemma 19.** For  $h \geq 0$  we have

$$(13) \quad y_h^x = \sum_{\Lambda \in \mathcal{PRT}^{\leq h}} \text{Cont}^x(\Lambda)$$

In particular,

$$(14) \quad \text{crit}(\tau^x) = \sum_{\Lambda \in \mathcal{PRT}} \text{Cont}^x(\Lambda),$$

where the right hand side converges for all  $x \in U$ .

*Proof.* We show the result holds by induction on  $h$ . For  $h = 0$   $y_0^x = 0$  which by convention is equal to the empty sum. Now suppose we have established Equation (13) for some  $h$ , let us show it holds for  $h + 1$ .

By Equation (12) and Equation (13) for  $h$  we have:

$$(15) \quad y_{h+1}^x = g^x(y_h^x) = \sum_{l \geq 0} \frac{(-1)}{l!} S_l^x((y_h^x)^{\otimes l}) =$$

$$(16) \quad = \sum_{l \geq 0} \frac{(-1)}{l!} S_l \left( \left( \sum_{a \in \mathcal{PRT}^{\leq h}} \text{Cont}^x(a) \right)^{\otimes l} \right)$$

On the other hand we have

$$(17) \quad \begin{aligned} \sum_{\Gamma \in \mathcal{PRT}^{\leq (h+1)}} A^x(\Gamma) &= \sum_{l \geq 0, (a_1, \dots, a_l) \in (\mathcal{PRT}^{\leq h})^l} \frac{(-1)}{l!} S_l(\bigotimes A^x(a_i)) \\ &= \sum_{l \geq 0} \frac{(-1)}{l!} S_l \left( \left( \sum_{a \in \mathcal{PRT}^{\leq h}} A^x(a) \right)^{\otimes l} \right) \end{aligned}$$

which shows that Equation (13) holds for  $h + 1$ , and the proof of the first claim is complete. The second claim, Eq (14), immediately follows.  $\square$

### 2.1. Coordinate Expression for the Perturbed Critical Point.

**Definition 20.** Let  $L$  be a finite set. An  $L$ -labeled rooted tree (or labeled-RT, if  $L$  is clear from the context) is a rooted tree  $\Gamma$  such that all the edges of  $\Gamma$  are labeled by elements of  $L$ . We will say an  $L$ -labeled rooted tree  $\Gamma$  has *output*  $l_0 \in L$  if the edge connecting the distinguished leaf  $v_{out}$  to the root is labeled  $l_0$ . The automorphism group  $\text{Aut}(\Gamma)$  of  $\Gamma$  consists of maps  $\psi : V(\Gamma) \rightarrow V(\Gamma)$  that fix the distinguished leaf

$v_{out}$  and such that  $u, v \in V(\Gamma)$  are connected by an edge labeled  $l$  iff  $\psi(u), \psi(v)$  are connected by an edge labeled  $l$ .

Let  $L$  be a set of size  $m = \dim V$ . In this subsection, we assume that  $B \in \text{Sym}^2(V^*)$  is the complexification of a positive definite pairing. In other words, there exists a basis  $\{e_i\}_{i \in L}$  to  $V$  such that

$$(18) \quad B = \sum_{i \in L} e_i^* \otimes e_i^*,$$

where  $\{e_i^*\}$  be the basis of  $V^*$  dual to  $\{e_i\}$ .

Define  $\theta_{i_1, \dots, i_l} : \mathbb{C}^r \rightarrow \mathbb{C}$  by

$$T_l^x = \sum_{i_1, \dots, i_l \in L} \theta_{i_1, \dots, i_l}(x) e_{i_1}^* \otimes \dots \otimes e_{i_l}^*.$$

Note that  $\theta_{i_1, \dots, i_l}(x)$  is invariant under any permutation of the indices  $i_1, \dots, i_l$ .

**Definition 21.** Let  $\Gamma$  be an  $L$ -labeled tree. For any internal vertex  $v$  such that the edges incident to  $v$  are labeled  $i_1, \dots, i_d$  write

$$C_v^x = -\theta_{i_1, \dots, i_d}(x).$$

We define the *amplitude function*  $A_\Gamma^x : \mathbb{C}^r \rightarrow \mathbb{C}$  of  $\Gamma$  by

$$(19) \quad A_\Gamma^x = \prod_{v \neq v_{out}} C_v^x,$$

where the product is taken over all vertices except  $v_{out}$ .

Write

$$\text{crit}(\tau^x) = \sum_{i \in L} \text{crit}(\tau^x)_i e_i$$

**Theorem 22.** For  $i \in L$  We have

$$(20) \quad \text{crit}(\tau^x)_i = \sum_{\Gamma} \frac{A_\Gamma^x}{|Aut(\Gamma)|}$$

where  $\Gamma$  ranges over a set of representatives for the isomorphism types of  $L$ -labeled rooted trees with output  $i$ .

In the special case when  $T_2^x = 0$  for all  $x$  the sum is taken over trees where all vertices are either leaves or of valency at least 3.

The following definition will be useful for proving the theorem.

**Definition 23.** (a) An  $L$ -labeled planar rooted tree  $\tilde{\Gamma}$  is a planar rooted tree together with a labeling.

(b) The *small amplitude*  $\tilde{A}_\Gamma^x$  of an  $L$ -labeled planar rooted tree  $\Gamma$  is defined by

$$\tilde{A}_\Gamma^x = \prod_{v \neq v_{out}} \frac{C_v^x}{(\deg(v) - 1)!},$$

where the product is taken over all vertices except  $v_{out}$ , and  $\deg(v)$  is the degree of  $v$ .

*Proof of Theorem 22.* The  $i^{th}$  coordinate of the expression (14) for  $\text{crit}(\tau^x)$  is

$$\text{crit}(\tau^x)_i = \sum_{\tilde{\Gamma}} \tilde{A}_{\tilde{\Gamma}}^x$$

where  $\tilde{\Gamma}$  ranges over the  $L$ -labeled planar rooted trees with output  $i$ . There is a forgetful map  $For$  from the set of labeled planar rooted trees to the set of labeled rooted trees, obtained by forgetting the cyclic orders.

Let  $\Gamma$  be a labeled rooted tree. We claim that size of the fiber over  $\Gamma$  of the forgetful map is given by

$$|For^{-1}(\{\Gamma\})| = \frac{1}{|Aut(\Gamma)|} \prod (\deg(v) - 1)!$$

where the product is taken over all vertices.

Indeed, fix some reference  $\tilde{\Gamma}_0 \in For^{-1}(\Gamma)$ . There is a group  $G$  of order  $|G| = \prod_{v \in V(\tilde{\Gamma}_0)} (\deg(v) - 1)!$  which acts transitively on  $For^{-1}(\Gamma)$  by changing the order of the subtrees. In fact  $G$  can be constructed as the semi-direct product of the symmetric groups  $\{S_{\deg(v)-1}\}_{v \in V(\tilde{\Gamma}_0)}$ . The stabilizer of  $\tilde{\Gamma}_0$  is naturally identified with  $Aut(\Gamma)$ , so by the orbit-stabilizer theorem the size of the fiber is  $\frac{1}{|Aut(\Gamma)|} \prod (\deg(v) - 1)!$  as claimed.

The amplitude function  $\tilde{A}_{\tilde{\Gamma}}^x$  is independent of the specific  $\tilde{\Gamma} \in For^{-1}(\Gamma)$ . By summing over all  $For$ -preimages we get,

$$\sum_{\tilde{\Gamma} \in For^{-1}(\Gamma)} \tilde{A}_{\tilde{\Gamma}}^x = \frac{\prod (\deg(v) - 1)!}{|Aut(\Gamma)|} \prod C_v^x (\deg(v) - 1)! = A_{\Gamma}^x.$$

As claimed.

The last claim follows from the observation that  $C_v = 0$  for any bivalent vertex.  $\square$

### 3. PROOFS OF THE MAIN THEOREMS

Let  $(\Sigma, Q, \{r_i\}_{i=1}^k)$  be a normalized KL-problem. Let

$$\Delta' = \{p_{\sigma} \in \mathbb{R}^{\Sigma} \mid (\forall \sigma \ p_{\sigma} > 0) \wedge \sum_{\sigma} p_{\sigma} = 1\}$$

denote the *open* simplex of probability distributions. We are looking for the probability distribution  $P \in \Delta'$  that minimizes

$$D(P \| Q) = \sum_{\sigma \in \Sigma} p_{\sigma} \log \left( \frac{p_{\sigma}}{q_{\sigma}} \right)$$

subject to the constraints  $\mathbb{E}_P r_i = \rho_i$ . Using the assumption  $q_\sigma > 0$ , simple analysis shows that  $D(P||Q)$  tends to  $+\infty$  as we approach the boundary of  $\Delta'$ . To be precise, for every  $M \in \mathbb{R}$  there is a compact subset  $K \subset \Delta'$  such that  $D(P||Q) \geq M$  for all  $P \in \Delta' - K$ . It follows that  $D(P||Q)$  obtains a minimum in  $\Delta'$ , which is unique since  $D(P||Q)$  is a strictly convex function of  $P$ . Now apply Lagrange multipliers

$$(21) \quad D(P||Q) - \sum_{i=1}^k \lambda_i (\mathbb{E}_P(r_i) - \rho_i) = \sum_{\sigma} Q_{\sigma} \frac{p_{\sigma}}{Q_{\sigma}} \left( -\log \frac{p_{\sigma}}{Q_{\sigma}} - \sum_{i=1}^k \lambda_i r_i(\sigma) - \lambda_0 \right) + \sum_i \lambda_i \rho_i + \lambda_0.$$

Set  $x_{\sigma} = \frac{p_{\sigma}}{Q_{\sigma}}$ . By requiring the vanishing of the partial derivatives with respect to  $x_{\sigma}$ , we find that

$$x_{\sigma} = \exp \left( - \sum_{i=1}^k \lambda_i r_i(\sigma) - \lambda_0 - 1 \right).$$

Plugging this into Equation 21 we are looking for the minimum of

$$(22) \quad \hat{\tau}^{\rho}(\lambda) = \mathbb{E}_Q \exp \left( - \sum_{i=1}^k \lambda_i r_i(\sigma) - \lambda_0 - 1 \right) + \sum \lambda_i \rho_i + \lambda_0$$

We will compute  $\text{crit}(\hat{\tau}^{\rho})$ , the critical point of  $\hat{\tau}^{\rho}(\lambda)$ , in two ways. First, by applying the Feynman calculus directly, we will express  $\text{crit}(\hat{\tau}^{\rho})$  as a sum over moment trees, and thus prove Theorem 8. Second, we can solve  $\partial_{\lambda_0} T = 0$  for  $\lambda_0$  and only then apply the Feynman calculus. This way will lead to a sum over cumulant trees, and the formula of Theorem 13.

### 3.1. Proof of moment tree theorems.

*Proof of Theorem 8.* Write  $\lambda'_i = \lambda_i$  for  $1 \leq i \leq k$  and  $\lambda'_0 = \lambda_0 + 1$ . Define  $\tau^{\rho}(\lambda')$  by analytically continuing  $\hat{\tau}^{\rho}(\lambda'_0 - 1, \lambda'_1, \dots, \lambda'_k)$  to  $\mathbb{C}^k \times V$  for  $V = \mathbb{C}^{k+1}$ . We assume that for  $0 \leq i \leq k$ ,  $\lambda'_i \in V^*$  is the dual to the standard basis  $\{e_i\}_{i=0}^k$  for  $V$ . Since we assume  $(\Sigma, Q, \{r_i\})$  is a normalized KL problem, the quadratic term of  $\tau^{\rho}$  is

$$\frac{1}{2} \sum_{0 \leq i, j \leq k} (\mathbb{E}_Q r_i r_j) \lambda'_i \lambda'_j = \sum_{i=0}^k \lambda'^2_i$$

and Equation (18) holds. Now apply Theorem 22. In the notation of Subsection 2.1, with  $x = \rho$  and  $y = \lambda'$ , we have

$$(23) \quad \theta^{\rho}_{i_1, \dots, i_l} = \begin{cases} (-1)^l \mathbb{E}_Q (\prod_{j=1}^l r_{i_j}) & \text{if } l \geq 3 \\ \rho_{i_1} & \text{if } l = 1 \text{ and } 1 \leq i_1 \leq k \\ 0 & \text{if } l = 1 \text{ and } i_1 = 0, \text{ or } l = 2. \end{cases}$$

and

$$(24) \quad \lambda'_i = \text{crit}(\tau^\rho)_i = \sum_{\Gamma} \frac{A_\Gamma^\rho}{|Aut(\Gamma)|},$$

where  $\Gamma$  ranges over a set of representatives for the isomorphism types of  $\{0, 1, \dots, k\}$ -labeled rooted trees, with output  $i$ , each vertex which is not a leaf is of valency at least 3, and, the edge of no leaf, other than the root, can be labeled 0. The last condition is a consequence of  $\theta_0^\rho = 0$ . Hence,  $\Gamma$  ranges over a set of representatives for the isomorphism types of  $i$ -moment trees as defined in 1.2.  $A_\Gamma^\rho$  is the amplitude defined in Equation 19. But by Equation 23,  $A_\Gamma^\rho$  is just  $A_\Gamma \rho^{I_\Gamma}$ , where  $A_\Gamma$  is the amplitude defined in Definition 7. Theorem 8 is thus proved.  $\square$

**Lemma 24.** The unique solution to  $\frac{\partial}{\partial \lambda'_0} \tau^\rho = 0$  is given by

$$(25) \quad \exp(\lambda'_0) = \sum_{\sigma \in \Sigma} q_\sigma \exp\left(-\sum_{i=1}^k \lambda_i r_i(\sigma)\right).$$

*Proof.* By Equation 22  $\tau^\rho(\lambda')$  has the form

$$(26) \quad \tau^\rho(\lambda') = \mathbb{E}_Q \exp\left(-\sum_{i=1}^k \lambda_i r_i(\sigma) - \lambda'_0\right) + \sum \lambda_i \rho_i + \lambda'_0 - 1$$

A direct computation gives  $\mathbb{E}_Q \exp\left(-\sum_{i=1}^k \lambda_i r_i(\sigma) - \lambda'_0\right) = 1$ , or

$$\mathbb{E}_Q \exp\left(-\sum_{i=1}^k \lambda_i r_i(\sigma)\right) = e^{\lambda'_0},$$

as claimed.  $\square$

*Proof of Theorem 9.* We have

$$\mu(\lambda_1, \dots, \lambda_k) = \frac{\sum_{\sigma} s(\sigma) q_{\sigma} \exp\left(-\sum_{i=1}^k \lambda_i r_i(\sigma)\right)}{\sum_{\sigma} q_{\sigma} \exp\left(-\sum_{i=1}^k \lambda_i r_i(\sigma)\right)}$$

which by Lemma 24 can be rewritten as

$$(27) \quad \mu(\lambda_1, \dots, \lambda_k) = \sum_{\sigma} s(\sigma) q_{\sigma} \exp\left(-\lambda'_0 - \sum_{i=1}^k \lambda_i r_i(\sigma)\right) = \sum_{\sigma} s(\sigma) q_{\sigma} \exp\left(-\sum_{i=0}^k \lambda'_i r_i(\sigma)\right),$$

where  $r_0 = 1$ , as before. Expand  $\exp(x) = \sum_{a \geq 0} \frac{x^a}{a!}$  to obtain

$$(28) \quad \begin{aligned} \mu(\lambda_1, \dots, \lambda_k) &= \sum_{\sigma} q_{\sigma} s(\sigma) \sum_{i_0, i_1, \dots, i_k \geq 0} \prod \frac{\prod_{j=0}^k (-r_j(\sigma))^{i_j} \lambda'_j(\sigma)^{i_j}}{\prod_{j=0}^k i_j!} \\ &= \sum_{i_0, i_1, \dots, i_k \geq 0} \frac{(-1)^{\sum_{j=0}^k i_j} \mathbb{E}_Q \left( s \prod_{j=0}^k r_j^{i_j} \right) \prod_{j=0}^k \lambda'_j(\sigma)^{i_j}}{\prod_{j=0}^k i_j!}. \end{aligned}$$

We can rewrite the last equation as

$$(29) \quad \mu(\lambda_1, \dots, \lambda_k) = \sum_{\Lambda} \frac{C_v}{|Aut(\Lambda)|} \lambda'^{I_{\Lambda}}$$

where the sum is taken over labeled rooted trees with a single non-leaf vertex  $v$ , where the output edge is labeled by  $k+1$ , and the other edge-labels are taken from  $\{0, 1, \dots, k\}$ . The coupling value  $C_v$  is as in Definition 7. By Equation 24, we may write the critical value  $\lambda'_i$  as a sum over labeled rooted trees. Substituting this into Equation 29 can be interpreted as a sum over all possible ways of replacing the leaves of  $\Lambda$  whose edge is labeled by  $i \in \{0, \dots, k\}$  by  $i$ -moment trees. It follows that

$$(30) \quad \mu = \sum_{\Gamma} \frac{A_{\Gamma} \rho^{I_{\Gamma}}}{|Aut(\Gamma)|}$$

where the sum is taken over  $k+1$ -moment trees  $\Gamma$ . The  $1/|Aut(\Gamma)|$  coefficient is obtained from the orbit-stabilizer theorem, by considering the action of  $Aut(\Lambda) = \prod_{j=0}^k i_j!$  on the trees obtained after substitution.  $\square$

### 3.2. Proofs of cumulant tree theorems.

*Proof of Theorem 13.* Recall that the critical  $\lambda'_i$ , for  $0 \leq i \leq k$ , are those which minimize

$$(31) \quad \tau^{\rho}(\lambda') = \mathbb{E}_Q \exp \left( - \sum_{i=1}^k \lambda'_i r_i(\sigma) - \lambda'_0 \right) + \sum \lambda'_i \rho_i + \lambda'_0 - 1$$

of Equation 26.

In order to get an expression in terms of joint cumulants, substitute  $\lambda'_0$  such that  $\frac{\partial}{\partial \lambda'_0} \tau^{\rho} = 0$ .  $\log(\sum_{\sigma \in \Sigma} q_{\sigma} \exp(-\sum_{i=1}^k \lambda'_i r_i(\sigma)))$ , by Lemma 24. Denote by  $T(\lambda_1, \dots, \lambda_k)$  the result of this substitution (recall  $\lambda_i = \lambda'_i$  for  $i \neq 0$ ). Then

$$T(\lambda_1, \dots, \lambda_k) = e^{-\log E + 1} E + \phi + \log E - 1 = \log E + \phi.$$

where  $E = \mathbb{E}_Q \exp(-\sum_{i=1}^k \lambda_i r_i(\sigma))$  and  $\phi = \sum_{i=1}^k \lambda_i \rho_i$ . By definition,  $\log E$  is the generating function for the joint cumulants. More precisely,

$$(32) \quad \log E = \sum_n \sum_{1 \leq i_1, \dots, i_n \leq k} (-1)^n \frac{\kappa(r_{i_1}, \dots, r_{i_n})}{n!} \prod_{j=1}^n \lambda_{i_j}.$$

. We claim that within the convergence domain there exists a unique critical value of  $\lambda_i$ . Indeed,  $\log E + \phi$  is convex since  $E$  is a linear combination of exponents with non negative coefficients. By Theorem 22, this unique critical value is given by Eq (10). The shape conditions

hold since the propagator is the quadratic tensor in  $\log E + \phi$ , hence all vertices must either be of degree 1 or degree at least 3.  $\square$

*Proof of Theorem 14.* As in the proof of Theorem 9 the strategy will be to substitute the expressions of  $\lambda_i$  in terms of cumulants in

$$\mu(\lambda_1, \dots, \lambda_k) = \frac{\sum_{\sigma} s(\sigma) q_{\sigma} \exp\left(-\sum_{i=1}^k \lambda_i r_i(\sigma)\right)}{\sum_{\sigma} q_{\sigma} \exp\left(-\sum_{i=1}^k \lambda_i r_i(\sigma)\right)}$$

Introduce a new variable  $\alpha$  and write

$$F(\lambda, \alpha) = F(\lambda_1, \dots, \lambda_k, \alpha) = \log\left(\mathbb{E}_Q \exp\left(\alpha s - \sum_{i=1}^k \lambda_i r_i(\sigma)\right)\right).$$

Then

$$\mu(\lambda_1, \dots, \lambda_k) = \frac{\partial}{\partial \alpha} F(\lambda_1, \dots, \lambda_k, \alpha)|_{\alpha=0}.$$

Now by the definition of cumulants again,

$$F(\lambda, \alpha) = \sum_{1 \leq i_1, \dots, i_n \leq k+1} (-1)^{n-n_{k+1}} \frac{\kappa(r_{i_1}, \dots, r_{i_n})}{n!} \prod_{j=1}^n \lambda_{i_j},$$

where  $n_{k+1}$  is the number of indices  $i_j = k+1$  and  $\lambda_{k+1} = \alpha$ . Hence the partial derivative in  $\alpha = 0$  is just

$$\mu(\lambda) = \sum_{1 \leq i_1, \dots, i_n \leq k} (-1)^n \frac{\kappa(r_{k+1}, r_{i_1}, \dots, r_{i_n})}{n!} \prod_{j=1}^n \lambda_{i_j}.$$

As in the proof of Theorem 9,  $\mu$  can thus be written as

$$(33) \quad \mu(\lambda_1, \dots, \lambda_k) = \sum_{\Lambda} \frac{C_v}{|Aut(\Lambda)|} \lambda^{I_{\Lambda}}$$

where the summation is taken over labeled rooted trees with a single non-leaf vertex  $v$ , output  $k+1$ , and the other labels are in  $\{1, \dots, k\}$ . The coupling value is as in Definition 12. The proof now ends by substituting the expressions for  $\lambda_i$ ,  $1 \leq i \leq k$  in terms of cumulants, just as in the proof of Theorem 9.  $\square$

## APPENDIX A. PROOF OF LEMMA 3

Recall that by Lemma 24

$$\log\left(\sum_{\sigma} q_{\sigma} \exp\left(-\sum_{i=1}^k \lambda_i r_i(\sigma)\right)\right) = 1 + \lambda_0 = 1 + \lambda_0(\lambda_1, \dots, \lambda_k).$$

Hence  $\lambda_0$  is an analytic function of  $\lambda_1, \dots, \lambda_k$  around  $\lambda_1 = \dots = \lambda_k = 0$ . Now,

$$\rho_i(\lambda_1, \dots, \lambda_k) = \sum_{\sigma} r_i(\sigma) q_{\sigma} \exp\left(-1 - \lambda_0(\lambda_1, \dots, \lambda_k) - \sum_{l=1}^k \lambda_l r_l(\sigma)\right) =$$



$$= \frac{\sum_{\sigma} r_i(\sigma) q_{\sigma} \exp\left(-\sum_{l=1}^k \lambda_l r_l(\sigma)\right)}{\exp(1 + \lambda_0(\lambda_1, \dots, \lambda_k))} = \frac{\sum_{\sigma} r_i(\sigma) q_{\sigma} \exp\left(-\sum_{l=1}^k \lambda_l r_l(\sigma)\right)}{\sum_{\sigma} q_{\sigma} \exp\left(-\sum_{l=1}^k \lambda_l r_l(\sigma)\right)}$$

So that  $\rho_i(\lambda_1, \dots, \lambda_k)$  is an analytic function of  $\lambda_1, \dots, \lambda_k$

The proof that  $\lambda_i = \lambda_i(\rho_1, \dots, \rho_k)$  is an analytic function of  $\rho_1, \dots, \rho_k$  around

$$\lambda_1 = \dots = \lambda_k = \rho_1 = \dots = \rho_k = 0,$$

uses the analytic inverse function theorem. It is enough to show that the Jacobian  $\frac{\partial(\rho_1, \dots, \rho_k)}{\partial(\lambda_1, \dots, \lambda_k)}$  is invertible for  $\lambda_1 = \dots = \lambda_k = 0$ .

But

$$(34) \quad \frac{\partial \rho_i}{\partial \lambda_j} = \frac{\left(\sum_{\sigma} r_i(\sigma) r_j(\sigma) q_{\sigma} \exp\left(-\sum_{l=1}^k \lambda_l r_l(\sigma)\right)\right) \left(\sum_{\sigma} q_{\sigma} \exp\left(-\sum_{l=1}^k \lambda_l r_l(\sigma)\right)\right)}{\left(\sum_{\sigma} q_{\sigma} \exp\left(-\sum_{l=1}^k \lambda_l r_l(\sigma)\right)\right)^2} - \frac{\left(\sum_{\sigma} r_i(\sigma) q_{\sigma} \exp\left(-\sum_{l=1}^k \lambda_l r_l(\sigma)\right)\right) \left(\sum_{\sigma} r_j(\sigma) q_{\sigma} \exp\left(-\sum_{l=1}^k \lambda_l r_l(\sigma)\right)\right)}{\left(\sum_{\sigma} q_{\sigma} \exp\left(-\sum_{l=1}^k \lambda_l r_l(\sigma)\right)\right)^2}.$$

Evaluation at  $\lambda_1 = \dots = \lambda_k = 0$  gives

$$\begin{aligned} \frac{\partial \rho_i}{\partial \lambda_j} \Big|_{\lambda_1 = \dots = \lambda_k = 0} &= \frac{(\sum_{\sigma} r_i(\sigma) r_j(\sigma) q_{\sigma}) (\sum_{\sigma} q_{\sigma}) - (\sum_{\sigma} r_i(\sigma) q_{\sigma}) (\sum_{\sigma} r_j(\sigma) q_{\sigma})}{(\sum_{\sigma} q_{\sigma})^2} = \\ &= \frac{\mathbb{E}_Q(r_i r_j) \cdot 1 - \mathbb{E}_Q(r_i) \mathbb{E}_Q(r_j)}{1^2} = \text{Cov}_Q(r_i, r_j). \end{aligned}$$

By assumption the KL constraint problem is normalized, hence

$$\frac{\partial \rho_i}{\partial \lambda_j} \Big|_{\lambda_1 = \dots = \lambda_k = 0} = \text{Cov}_Q(r_i, r_j) = \delta_{i,j}.$$

The Jacobian  $\frac{\partial(\rho_1, \dots, \rho_k)}{\partial(\lambda_1, \dots, \lambda_k)} = I_k$  is thus invertible.  $\square$

## REFERENCES

- [Ja57I] E. T. Jaynes *Information theory and statistical mechanics*. Phys. Rev. 106 (1957), issue 4, 620-630. American Physical society.
- [Ja57II] E. T. Jaynes *Information theory and statistical mechanics II*. Phys. Rev. 108 (1957), issue 2, 171-190. American Physical society.
- [Et02] P. Etingof *Geometry and Quantum Field Theory*. MIT OpenCourseware 18.238 (2002), Ch. 2,3.

DEPARTMENT OF MATHEMATICS, THE HEBREW UNIVERSITY OF JERUSALEM,  
JERUSALEM, ISRAEL

*E-mail address:* `tomer.schlank@gmail.com`

INSTITUTE FOR THEORETICAL STUDIES, ETH ZÜRICH  
ZÜRICH, SWITZERLAND

*E-mail address:* `ran.tessler@mail.huji.ac.il`

DEPARTMENT OF MATHEMATICS, THE HEBREW UNIVERSITY OF JERUSALEM,  
JERUSALEM, ISRAEL

*E-mail address:* `amitai.zernik@mail.huji.ac.il`